



US006066454A

United States Patent [19]**Lipshutz et al.**[11] **Patent Number:** **6,066,454**[45] **Date of Patent:** ***May 23, 2000**[54] **COMPUTER-AIDED PROBABILITY BASE
CALLING FOR ARRAYS OF NUCLEIC ACID
PROBES ON CHIPS**

5,470,710	11/1995	Weiss et al.	435/6
5,502,773	3/1996	Tibbetts et al.	382/129
5,503,980	4/1996	Cantor	435/6

[75] **Inventors:** **Robert J. Lipshutz**, Palo Alto; **Michael G. Walker**, Sunnyvale, both of Calif.[73] **Assignee:** **Affymetrix, Inc.**, Santa Clara, Calif.[*] **Notice:** This patent is subject to a terminal disclaimer.[21] **Appl. No.:** **08/948,896**[22] **Filed:** **Oct. 10, 1997****Related U.S. Application Data**

[63] Continuation of application No. 08/528,656, Sep. 14, 1995, Pat. No. 5,733,729.

[51] **Int. Cl.⁷** **C12Q 1/68; C07H 21/04; G06K 9/18**[52] **U.S. Cl.** **435/6; 436/497; 536/23.1; 536/24.32; 382/129**[58] **Field of Search** **435/6, 973; 536/23.1, 536/24.3, 24.31, 24.32; 935/78; 436/497; 382/129**[56] **References Cited****U.S. PATENT DOCUMENTS**

5,002,867	3/1991	Macevitz	435/6
5,143,854	9/1992	Pirung et al.	436/518
5,202,231	4/1993	Drmanac et al.	435/6
5,235,626	8/1993	Flamholz et al.	378/34
5,288,514	2/1994	Ellman	427/2
5,365,455	11/1994	Tibbetts et al.	364/497
5,384,261	1/1995	Winkler et al.	436/518
5,445,934	8/1995	Fodor et al.	435/6

FOREIGN PATENT DOCUMENTS

WO 89/10977	11/1989	WIPO
WO 92/10092	6/1992	WIPO
WO 92/10588	6/1992	WIPO
WO 95/11995	5/1995	WIPO

OTHER PUBLICATIONS

Fodor et al., "Light-Directed Spatially Addressable Parallel Chemical Synthesis," Science, vol. 251, Feb. 15, 1991, pp. 767-773.

Brown et al., An Inexpensive MSI/LSI Mask Making System, Proceedings of 1981 Univ. Govt. Indus. Microelec. Symposium, May 26-27, 1981, pp. III-31 through III-38.

Dear et al., "A Sequence Assembly And Editing Program For Efficient Management of Large Projects," Nucleic Acids Research, vol. 19, No. 14, 1991 Oxford Univ. Press, pp. 3907-3911.

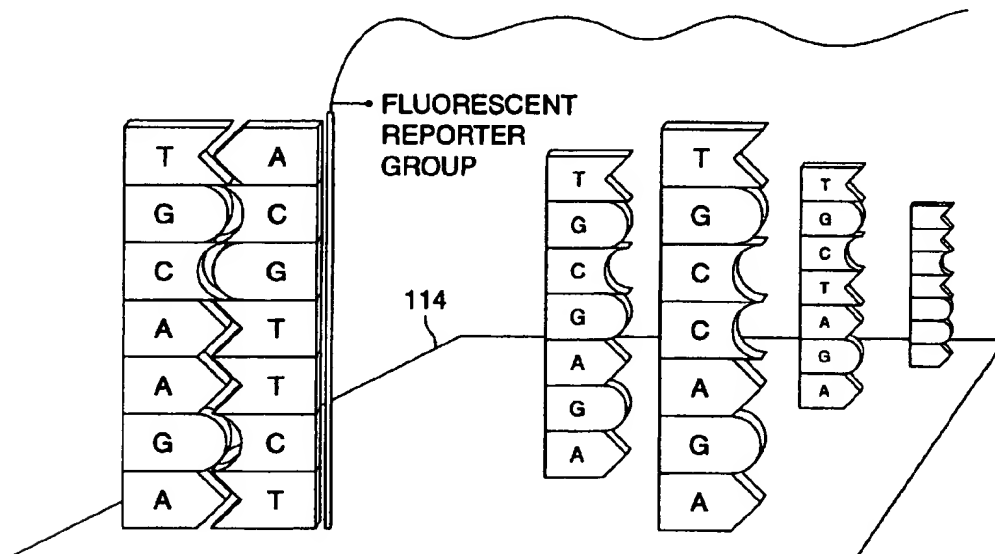
R. Drmanac et al., "Journal of Biomolecular Structure & Dynamics," 8(5): 1085-1102, 1991.

Primary Examiner—Stephanie Zitomer*Assistant Examiner*—Betty J Forman*Attorney, Agent, or Firm*—Ritter, Van Pelt & Yi

[57]

ABSTRACT

A computer system for analyzing nucleic acid sequences is provided. The computer system is used to calculate probabilities for determining unknown bases by analyzing the fluorescence intensities of hybridized nucleic acid probes on biological chips. Additionally, information from multiple experiments is utilized to improve the accuracy of calling unknown bases.

38 Claims, 8 Drawing Sheets

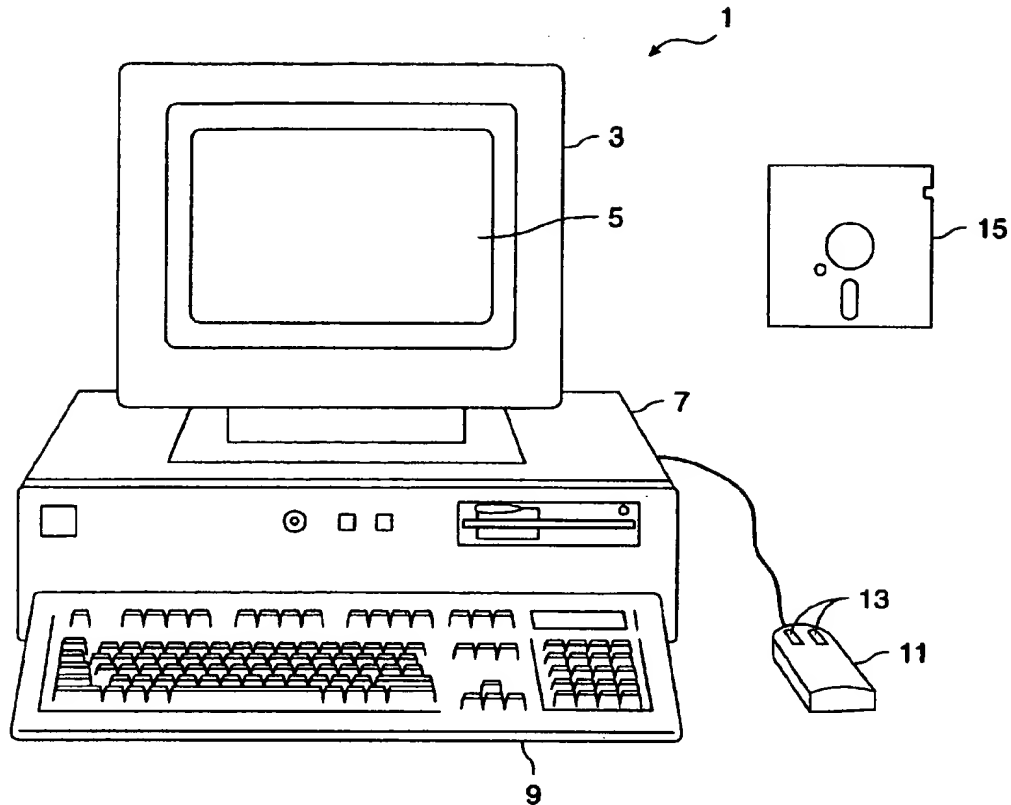


FIG. 1

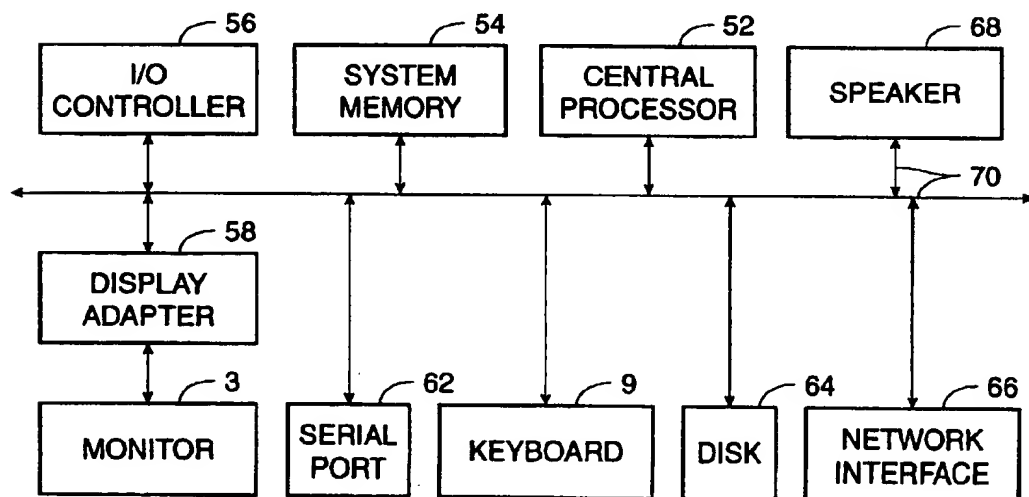


FIG. 2

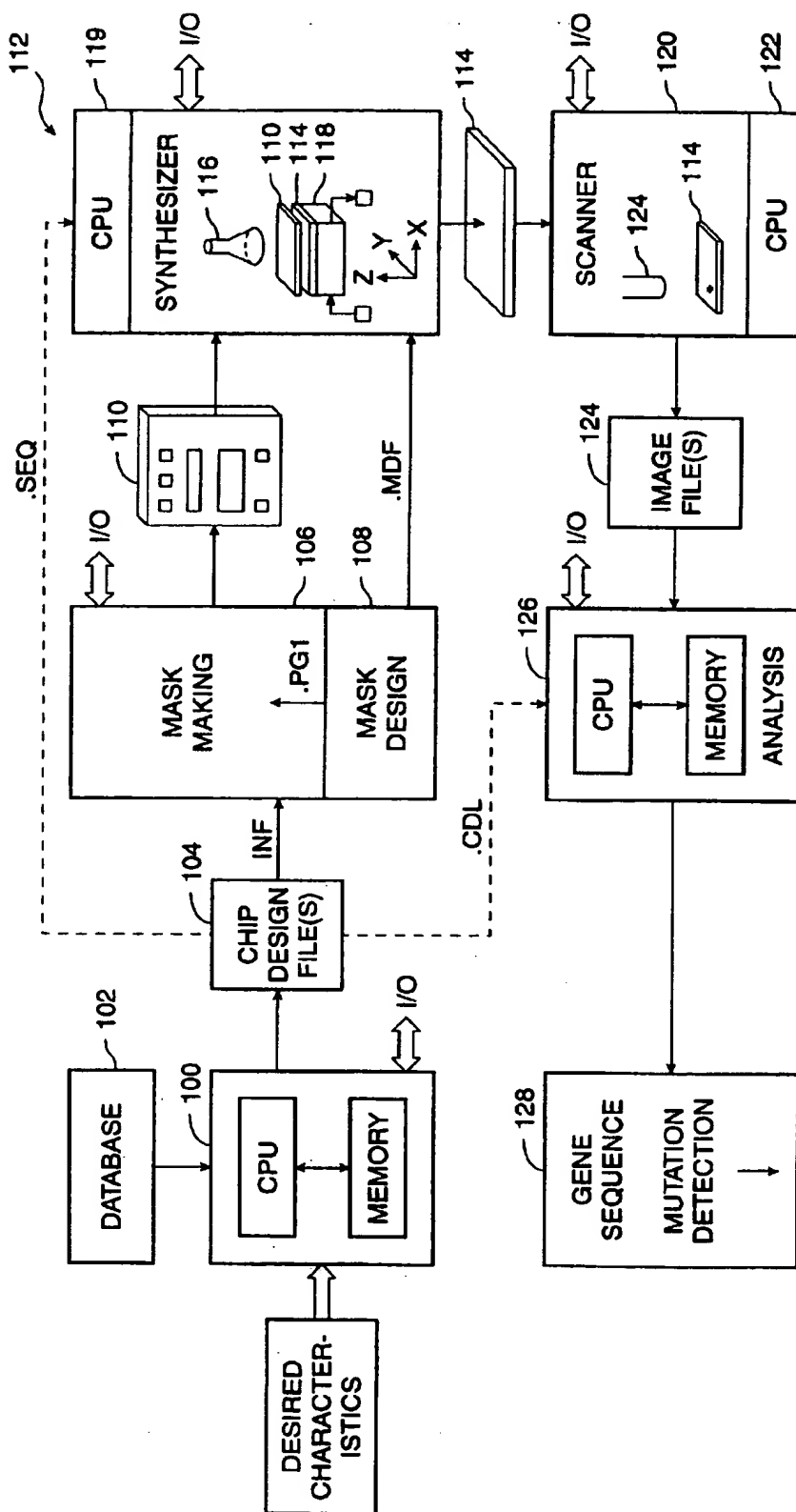


FIG. 3

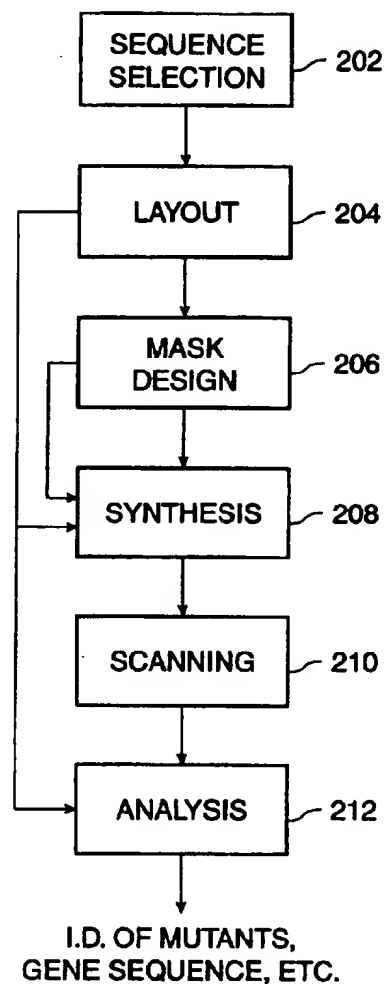


FIG. 4

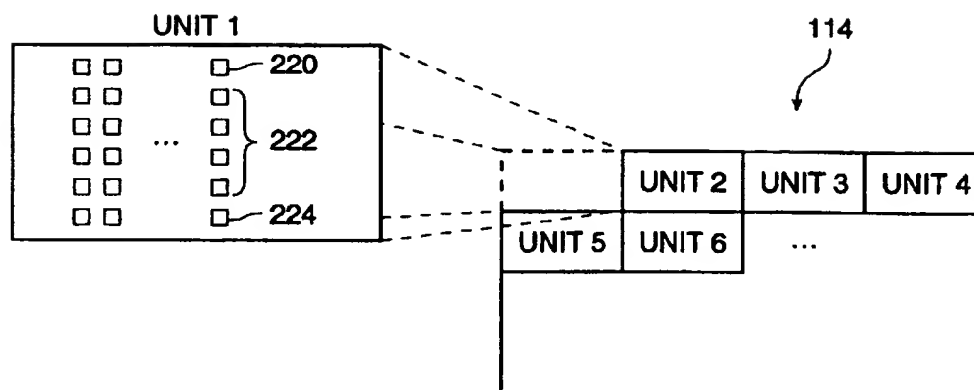


FIG. 5

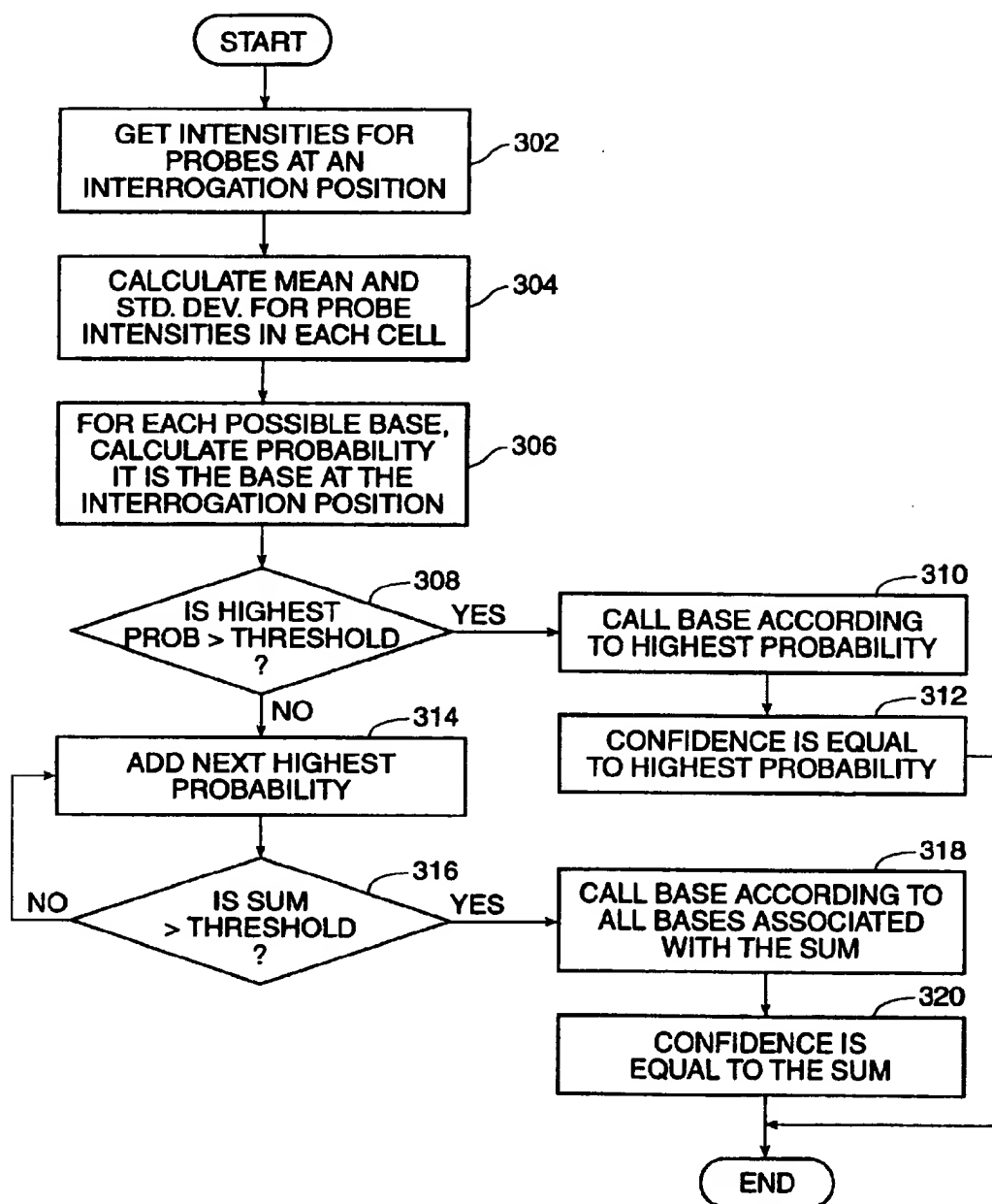
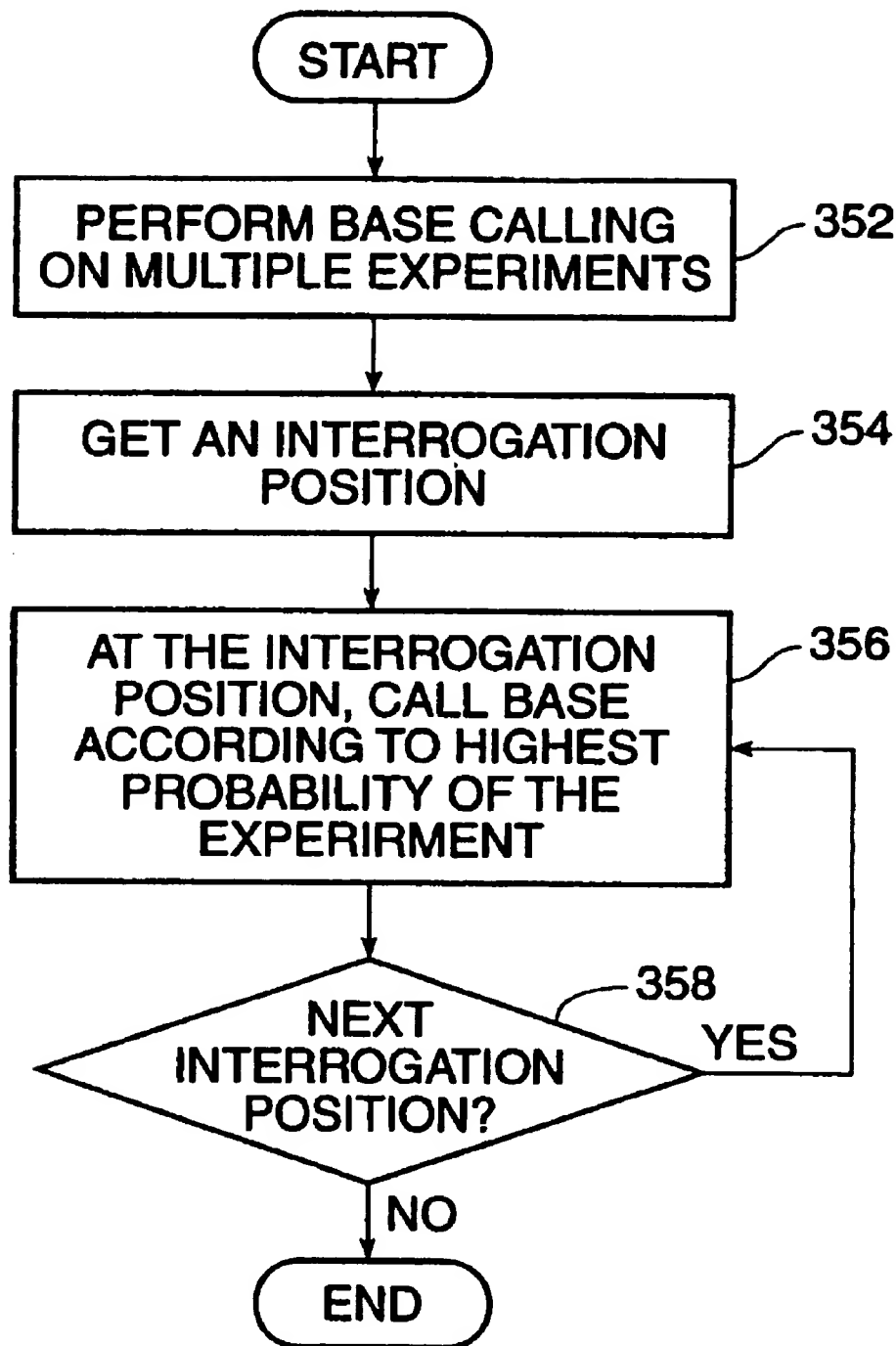
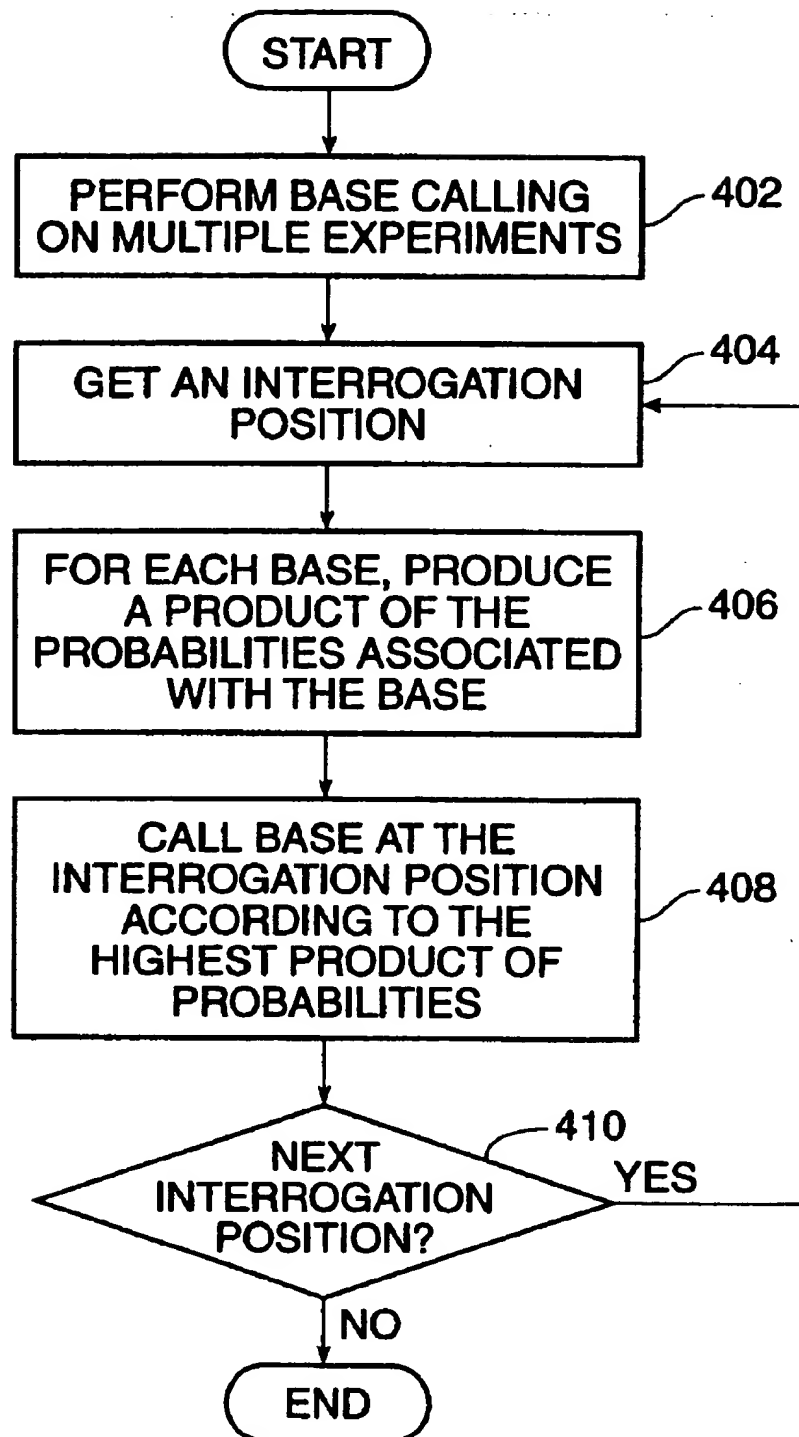
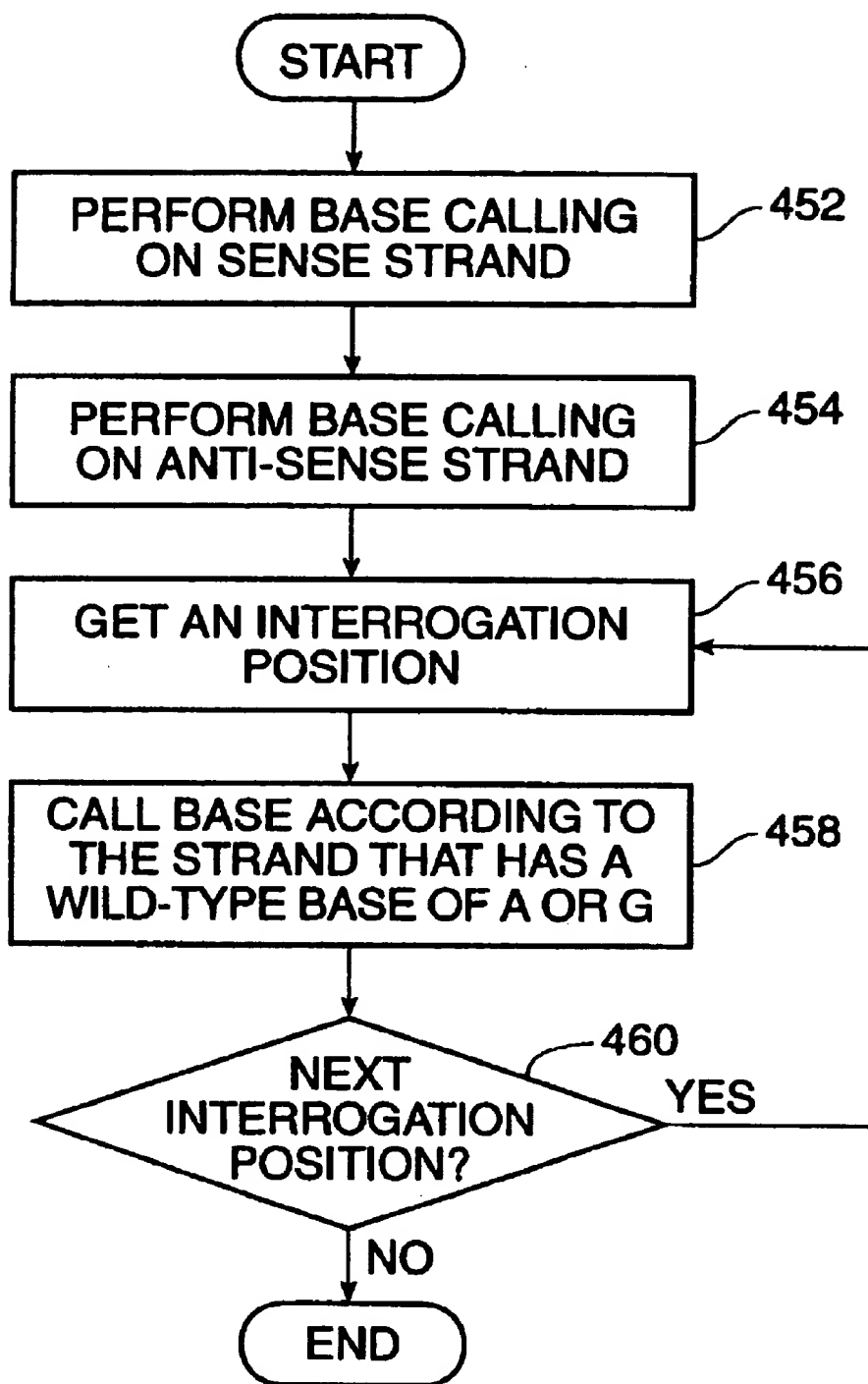


FIG. 9

**FIG. 10**

**FIG. 11**

**FIG. 12**

COMPUTER-AIDED PROBABILITY BASE CALLING FOR ARRAYS OF NUCLEIC ACID PROBES ON CHIPS

This is a continuation application of prior application 5
Ser. No. 08/528,656 filed on Sep. 14, 1995, now U.S. Pat.
No. 5,733,729, the disclosure of which is incorporated
herein by reference.

GOVERNMENT RIGHTS NOTICE

Portions of the material in this specification arose under
the cooperative agreement 70NANB5H1031 between
Affymetrix, Inc. and the Department of Commerce through
the National Institute of Standards and Technology.

COPYRIGHT NOTICE

A portion of the disclosure of this patent document
contains material which is subject to copyright protection.
The copyright owner has no objection to the xerographic
reproduction by anyone of the patent document or the patent
disclosure in exactly the form it appears in the Patent and
Trademark Office patent file or records, but otherwise
reserves all copyright rights whatsoever.

SOFTWARE APPENDIX

A Software Appendix comprising twenty one (21) sheets 25
is included herewith.

BACKGROUND OF THE INVENTION

The present invention relates to the field of computer
systems. More specifically, the present invention relates to
computer systems for evaluating and comparing biological
sequences.

Devices and computer systems for forming and using
arrays of materials on a substrate are known. For example,
PCT application WO92/10588, incorporated herein by ref-
erence for all purposes, describes techniques for sequencing
or sequence checking nucleic acids and other materials.
Arrays for performing these operations may be formed in
arrays according to the methods of, for example, the pio-
neering techniques disclosed in U.S. Pat. No. 5,143,854 and
U.S. patent application Ser. No. 08/249,188, now U.S. Pat.
No. 5,571,639, both incorporated herein by reference for all
purposes.

According to one aspect of the techniques described
therein, an array of nucleic acid probes is fabricated at
known locations on a chip or substrate. A fluorescently
labeled nucleic acid is then brought into contact with the
chip and a scanner generates an image file (also called a cell
file) indicating the locations where the labeled nucleic acids
bound to the chip. Based upon the image file and identities
of the probes at specific locations, it becomes possible to
extract information such as the monomer sequence of DNA
or RNA. Such systems have been used to form, for example,
arrays of DNA that may be used to study and detect
mutations relevant to cystic fibrosis, the P53 gene (relevant
to certain cancers), HIV, and other genetic characteristics.

Innovative computer-aided techniques for base calling are
disclosed in U.S. patent application Ser. No. 08/327,525,
which is incorporated by reference for all purposes.
However, improved computer systems and methods are still
needed to evaluate, analyze, and process the vast amount of
information now used and made available by these pioneer-
ing technologies.

SUMMARY OF THE INVENTION

An improved computer-aided system for calling unknown
bases in sample nucleic acid sequences from multiple

nucleic acid probe intensities is disclosed. The present
invention is able to call bases with extremely high accuracy
(up to 98.5%). At the same time, confidence information
may be provided that indicates the likelihood that the base
has been called correctly. The methods of the present
invention are robust and uniformly optimal regardless of the
experimental conditions.

According to one aspect of the invention, a computer
system is used to identify an unknown base in a sample
nucleic acid sequence by the steps of: inputting a plurality of
hybridization probe intensities, each of the probe intensities
corresponding to a nucleic acid probe; for each of the
plurality of probe intensities, determining a probability that
the corresponding nucleic acid probe best hybridizes with
the sample nucleic acid sequence; and calling the unknown
base according to the nucleic acid probe with the highest
associated probability.

According to another aspect of the invention, an unknown
base in a sample nucleic acid sequence is called by a base
call with the highest probability of correctly calling the
unknown base. The unknown base in the sample nucleic acid
sequence is identified by the steps of: inputting multiple base
calls for the unknown base, each of the base calls having an
associated probability which represents a confidence that the
unknown base is called correctly; selecting a base call that
has a highest associated probability; and calling the
unknown base according to the selected base call. The
multiple base calls are typically produced from multiple
experiments. The multiple experiments may be performed
on the same chip utilizing different parameters (e.g., nucleic
acid probe length).

According to yet another aspect of the invention, an
unknown base in a sample nucleic acid sequence is called
according to multiple base calls that collectively have the
highest probability of correctly calling the unknown base.
The unknown base in the sample nucleic acid sequence is
identified by the steps of: inputting multiple probabilities for
each possible base for the unknown base, each of the
probabilities representing a probability that the unknown
base is an associated base; producing a product of probabili-
ties for each possible base, each product being associated
with a possible base; and calling the unknown base accord-
ing to a base associated with a highest product. The multiple
base calls are typically produced from multiple experiments.
The multiple experiments may be performed on the same
chip utilizing different parameters (e.g., nucleic acid probe
length).

According to another aspect of the invention, both strands
of a DNA molecule are analyzed to increase the accuracy of
identifying an unknown base in a sample nucleic acid
sequence by the steps of: inputting a first base call for the
unknown base, the first base call determined from a first
nucleic acid probe that is equivalent to a portion of the
sample nucleic acid sequence including the unknown base;
inputting a second base call for the unknown base, the
second base call determined from a second nucleic acid
probe that is complementary to a portion of the sample
nucleic acid sequence including the unknown base; selecting
one of the first or second nucleic acid probes that has a base
at an interrogation position which has a high probability of
producing correct base calls; and calling the unknown base
according to the selected one of the first or second nucleic
acid probes.

A further understanding of the nature and advantages of
the inventions herein may be realized by reference to the
remaining portions of the specification and the attached
drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 illustrates an example of a computer system used to execute the software of the present invention;

FIG. 2 shows a system block diagram of a typical computer system used to execute the software of the present invention;

FIG. 3 illustrates an overall system for forming and analyzing arrays of biological materials such as DNA or RNA;

FIG. 4 is an illustration of the software for the overall system;

FIG. 5 illustrates the global layout of a chip formed in the overall system;

FIG. 6 illustrates conceptually the binding of probes on chips;

FIG. 7 illustrates probes arranged in lanes on a chip;

FIG. 8 illustrates a hybridization pattern of a target on a chip with a reference sequence as in FIG. 7;

FIG. 9 illustrates the high level flow of the probability base calling method; and

FIG. 10 illustrates the flow of the maximum probability method;

FIG. 11 illustrates the flow of the product of probabilities method; and

FIG. 12 illustrates the flow of the wild-type base preference method.

DESCRIPTION OF THE PREFERRED EMBODIMENT

Contents

- I. General
- II. Probability Base Calling Method
- III. Maximum Probability Method
- IV. Product of Probabilities Method
- V. Wild-Type Base Preference Method
- VI. Software Appendix

I. General

In the description that follows, the present invention will be described in reference to a Sun Workstation in a UNIX environment. The present invention, however, is not limited to any particular hardware or operating system environment. Instead, those skilled in the art will find that the systems and methods of the present invention may be advantageously applied to a variety of systems, including IBM personal computers running MS-DOS or Microsoft Windows. Therefore, the following description of specific systems are for purposes of illustration and not limitation.

FIG. 1 illustrates an example of a computer system used to execute the software of the present invention. FIG. 1 shows a computer system 1 which includes a monitor 3, screen 5, cabinet 7, keyboard 9, and mouse 11. Mouse 11 may have one or more buttons such as mouse buttons 13. Cabinet 7 houses a floppy disk drive 14 and a hard drive (not shown) that may be utilized to store and retrieve software programs incorporating the present invention. Although a floppy disk 15 is shown as the removable media, other removable tangible media including CD-ROM and tape may be utilized. Cabinet 7 also houses familiar computer components (not shown) such as a processor, memory, and the like.

FIG. 2 shows a system block diagram of computer system 1 used to execute the software of the present invention. As

in FIG. 1, computer system 1 includes monitor 3 and keyboard 9. Computer system 1 further includes subsystems such as a central processor 52, system memory 54, I/O controller 56, display adapter 58, serial port 62, disk 64, network interface 66, and speaker 68. Other computer systems suitable for use with the present invention may include additional or fewer subsystems. For example, another computer system could include more than one processor 52 (i.e., a multi-processor system) or memory cache.

Arrows such as 70 represent the system bus architecture of computer system 1. However, these arrows are illustrative of any interconnection scheme serving to link the subsystems. For example, speaker 68 could be connected to the other subsystems through a port or have an internal direct connection to central processor 52. Computer system 1 shown in FIG. 2 is but an example of a computer system suitable for use with the present invention. Other configurations of subsystems suitable for use with the present invention will be readily apparent to one of ordinary skill in the art.

The VLSIPS technology provides methods of making very large arrays of oligonucleotide probes on very small chips. See U.S. Pat. No. 5,143,854 and PCT patent publication Nos. WO 90/15070 and 92/10092, each of which is incorporated by reference for all purposes. The oligonucleotide probes on the "DNA chip" are used to detect complementary nucleic acid sequences in a sample nucleic acid of interest (the "target" nucleic acid).

The present invention provides methods of analyzing hybridization intensity files for a chip containing hybridized nucleic acid probes. In a representative embodiment, the files represent fluorescence data from a biological array, but the files may also represent other data such as radioactive intensity data. Therefore, the present invention is not limited to analyzing fluorescent measurements of hybridizations but may be readily utilized to analyze other measurements of hybridization.

For purposes of illustration, the present invention is described as being part of a computer system that designs a chip mask, synthesizes the probes on the chip, labels the nucleic acids, and scans the hybridized nucleic acid probes. Such a system is fully described in U.S. patent application Ser. No. 08/249,188, now U.S. Pat. No. 5,591,639 which has been incorporated by reference for all purposes. However, the present invention may be used separately from the overall system for analyzing data generated by such systems.

FIG. 3 illustrates a computerized system for forming and analyzing arrays of biological materials such as RNA or DNA. A computer 100 is used to design arrays of biological polymers such as RNA or DNA. The computer 100 may be, for example, an appropriately programmed Sun Workstation or personal computer or workstation, such as an IBM PC equivalent, including appropriate memory and a CPU as shown in FIGS. 1 and 2. The computer system 100 obtains inputs from a user regarding characteristics of a gene of interest, and other inputs regarding the desired features of the array. Optionally, the computer system may obtain information regarding a specific genetic sequence of interest from an external or internal database 102 such as GenBank. The output of the computer system 100 is a set of chip design computer files 104 in the form of, for example, a switch matrix, as described in PCT application WO 92/10092, and other associated computer files.

The chip design files are provided to a system 106 that designs the lithographic masks used in the fabrication of